

Supplementary Material for: Learning spatio-temporal feature representations for video-based gaze estimation

Alexandre Personnic, Mihai Bâce

KU Leuven, Department of Computer Science, Group T Leuven Campus, Leuven, Belgium

{alexandre.personnic, mihai.bace}@kuleuven.be

Overview

This document provides supplementary details for our paper. We include the complete results of our ablation study, a more granular performance analysis on the EVE validation set, and a detailed breakdown of our model’s complexity.

1. Full Ablation Study Results

Table 1 presents the complete results of our ablation study on both the EVE validation and test sets. These results provide the empirical evidence for the design choices discussed in the main paper.

Table 1. Results of our ablation study of ST-Gaze’s architecture on the EVE *validation* and *test* set. Our full model provides the best performance, with the SAM and our recurrence structure being the most critical components for generalization to the test set.

| Model Configuration | Angular Error (°) | |
|--|-------------------|-------------|
| | Validation | Test |
| ST-Gaze (Full Model) | 1.86 | 2.58 |
| <i>Ablating Core Modules:</i> | | |
| w/o ECA Module | 2.03 | 2.56 |
| w/o Self-Attention Module | 2.02 | 4.84 |
| w/o GRU (Static Model) | 2.24 | 2.88 |
| Spatial Pooling before GRU | 2.00 | 2.79 |
| <i>Ablating Backbone & Fusion:</i> | | |
| ResNet-18 Backbone | 2.25 | 4.22 |
| EfficientNet-B0 Backbone | 2.10 | 2.78 |
| EfficientNet-B7 Backbone | 2.12 | 2.90 |
| EfficientNet-V2-S Backbone | 2.69 | 3.18 |
| Early Fusion Strategy | 2.26 | 3.26 |

The results in Table 1 confirm the trends discussed in the main paper, while also revealing a critical discrepancy between the validation and test sets. The validation set is a useful indicator for many design choices; for instance, it correctly shows that our spatio-temporal recur-

rence (1.86°) is superior to the conventional “Spatial Pooling before GRU” approach (2.00°). However, the validation set fails to predict the critical importance of certain components for robust generalization. Notably, removing the Self-Attention Module or using a ResNet-18 backbone results in only a modest performance drop on the validation set (to 2.02° and 2.25° respectively). In contrast, these same configurations catastrophically fail on the test set (4.84° and 4.22°). This discrepancy strongly suggests that the validation set, while useful for initial tuning, does not fully capture the challenging appearance and pose variations present in the test set, which are necessary to evaluate true “in-the-wild” generalization.

2. Detailed Performance Analysis on the EVE Validation Set

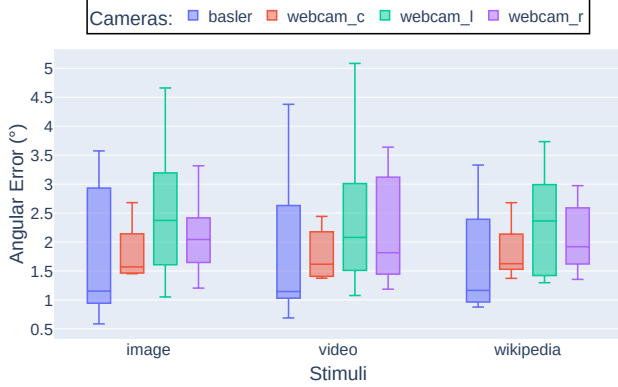
To provide a more granular understanding of our model’s behaviour, we present a detailed analysis of its performance on the EVE validation set in Figure 1.

2.1. Influence of Camera and Stimuli

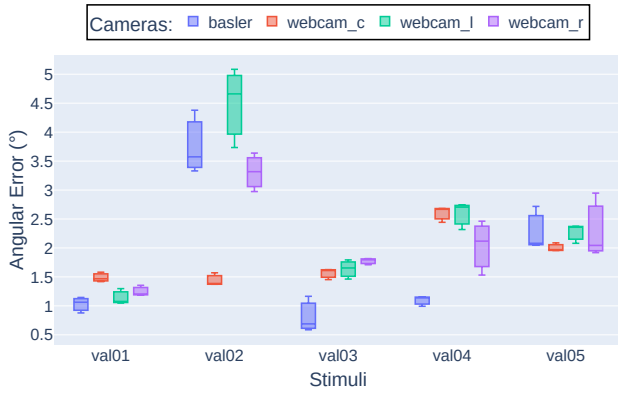
Panel 1a breaks down performance by camera view and stimulus type. The model is largely robust to the content being viewed, with only a slight, expected increase in error for dynamic *Video* stimuli. A more significant trend emerges across the camera views. While the high-quality, centrally-mounted ‘basler’ camera yields the lowest average error, the *webcam_c* offers the most consistent predictions with the lowest variance. The notable performance gap between the left and right webcams (*webcam_l* vs. *webcam_r*) suggests an underlying asymmetry in the dataset, likely from environmental factors such as uneven lighting, which architectural choices like flipping input images cannot fully compensate for.

2.2. Influence of Participant

While environmental variations are important, Panel 1b reveals that the most dominant source of performance variance is the individual participant. The model performs



(a) Performance across stimuli and camera views. The central webcam (*webcam_c*) offers the most stable predictions, while the asymmetry between *webcam_l* and *webcam_r* suggests a dataset bias.



(b) Performance across individual participants. Note the significant variance, with *val02* representing a challenging outlier, highlighting the impact of person-specific factors.

Figure 1. Detailed performance analysis of ST-Gaze on the EVE validation set. These figures provide a granular breakdown of the aggregate results presented in the main paper’s ablation study.

exceptionally well for participants *val01* and *val03*, achieving consistently low average errors, typically between 0.75° and 1.75° , with low deviation across all cameras. This demonstrates that our generalized model can achieve very high accuracy for certain individuals.

In stark contrast, participant *val02* is a notable outlier, with errors ranging from 3.0° to 5.0° on three of the four cameras. Interestingly, performance on the central *webcam_c* view is substantially better for this participant, suggesting a strong sensitivity to specific head poses or camera angles for certain individuals. The remaining participants exhibit more nuanced patterns. For *val04*, the model is highly accurate with the *basler* camera (1.0° error) but struggles more with the lower-resolution webcams ($2.0^\circ - 2.5^\circ$ error), indicating a sensitivity to input image quality. Participant *val05* shows consistent, moderate performance across all views, with an error between 2.0°

and 2.5° . This detailed per-participant analysis underscores the central challenge of person-independent gaze estimation and motivates the need for the person-specific adaptation methods discussed in the main paper.

3. Model Complexity

For completeness, we provide a detailed breakdown of our model’s complexity. Table 2 details the parameter distribution across the main components of ST-Gaze.

Table 2. Parameter distribution within our proposed ST-Gaze. The majority of parameters reside in the feature encoders, while our novel attention and recurrence modules are highly parameter-efficient.

| Component | Parameters | % of Total |
|-----------------------|-------------|---------------|
| Eye Encoder | 10 M | 47.37 |
| Face Encoder | 10 M | 47.19 |
| ECA Module | 5 | 0.00 |
| Self-Attention Module | 815 k | 3.80 |
| Spatio-Temporal GRU | 309 k | 1.44 |
| Gaze Regression | 41 k | 0.19 |
| Total | 21 M | 100.00 |

The novel spatio-temporal recurrence module that is central to our contribution comprises less than 2% of the total parameters, demonstrating its efficiency. Additionally, the number of parameters in the EfficientNet-B3 encoder (10 M) [35] is similar to ResNet18 (11 M) [16], as such, our usage of this encoder does not incur any significant increase in computational cost compared to other dual-input models like FE-NET [33] or STTDN [22].

4. Code Repository

A public version of our repository can be found in our institutional GitLab at the following address: <https://gitlab.kuleuven.be/u0172623/ST-Gaze>. It contains instructions to set-up the appropriate environment to train and evaluate our model as well the pretrained weights corresponding to our best result.